Quantifying the Probability of Existential Catastrophe: A Reply to Beard et al.

Seth D. Baum
Global Catastrophic Risk Institute

Futures, vol. 123 (October 2020), article 102608, DOI 10.1016/j.futures.2020.102608
This version 10 August 2020.

Abstract
A recent article by Beard, Rowe, and Fox (BRF) evaluates ten methodologies for quantifying the probability of existential catastrophe. This article builds on BRF's valuable contribution. First, this article describes the conceptual and mathematical relationship between the probability of existential catastrophe and the severity of events that could result in existential catastrophe. It discusses complications in this relationship arising from catastrophes occurring at different speeds and from multiple concurrent catastrophes. Second, this article revisits the ten BRF methodologies, finding an inverse relationship between a methodology's ease of use and the quality of results it produces—in other words, achieving a higher quality of analysis will in general require a larger investment in analysis. Third, the manuscript discusses the role of probability quantification in the management of existential risks, describing why the probability is only sometimes needed for decision-making and arguing that analyses should support real-world risk management decisions and not just be academic exercises. If the findings of this article are taken into account, together with BRF's evaluations of specific methodologies, then risk analyses of existential catastrophe may tend to be more successful at understanding and reducing the risks.

Keywords: existential risk; global catastrophic risk; probability; severity; risk analysis

1. Introduction

A recent article by Beard, Rowe, and Fox (Beard et al. 2020; henceforth BRF) surveys methods used to quantify the probabilities of global and existential catastrophes. BRF is a valuable contribution to the study of global catastrophic risk and existential risk. It documents the range of methods in use and evaluates their strengths and limitations, providing both a good resource for researchers wishing to get up to speed on the topic and constructive guidance for future work. In this article, I provide some further commentary on the quantification of the probability of global/existential catastrophe, making some points not made by BRF or other prior literature.

At the outset, I wish to express my wholehearted agreement with BRF in the importance of sound methodology for quantifying the probabilities (and severities) of global and existential catastrophes. I regret that I share the view that "within the nascent field of Existential Risk research people have been insufficiently discriminating" in the selection and use of quantification methods (BRF, Section 7). The following remarks join BRF in seeking to improve this situation.

2. Some Considerations Related to Severity

2.1 The Definition of Existential Catastrophe

BRF defines existential catastrophe as "the collapse of civilization or the extinction of humanity". This is a reasonable definition, but there are others. First, Bostrom (2002) defines existential catastrophe as an event that "would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential". This definition emphasizes long-term effects. The collapse of civilization would presumably have some long-term effects, though whether the effects would last into the distant future is a complex matter (Baum et al. 2019). Second, Tonn and Stiefel (2013) define existential catastrophe as human extinction, i.e. loss of the existence of the human species. This paper will use the BRF definition while noting the existence of other definitions for existential catastrophe, for related concepts involving large losses of value such as global catastrophe (Morrison 1992), ultimate harm (Persson and Savulescu 2012), and oblivion (Tonn 1999), and for other concepts involving large-scale harm such as astronomical suffering (Sotala and Gloor 2017).

2.2 Probability Vs. Severity

BRF focuses on methodology for quantifying the probability of existential catastrophe, with limited attention other factors including severity. (This focus was an explicit decision; see BRF, Section 1.) That is not unreasonable for a first pass at the topic—there is plenty to say about the probability methodology. However, it is important to recognize the interplay between probability and severity.

All quantifications of the probability of existential catastrophe require at least some attention to severity for the simple reason that existential catastrophe is a type of event defined by its severity. More precisely, existential catastrophe is generally defined in terms of some minimum severity threshold; in BRF, the minimum threshold is the collapse of civilization. To assess the probability of existential catastrophe is to assess the probability of a harm greater or equal to the collapse of civilization. In the terminology of catastrophe risk analysis, the probability of existential catastrophe is an "exceedance probability", i.e. the probability of harms exceeding some minimum threshold (Grossi et al. 2005).

Quantifying the probability of specific existential catastrophe events (such as a nuclear war or Earth-asteroid collision) requires additional attention to severity. The probability can be decomposed into two constituent parts as follows:

$$P_{EC}=\mathrm{P}_1 * P_2 \quad (1)$$

In Equation 1, P_{EC} is the probability of existential catastrophe from some event; P_1 is the probability of the initial catastrophe event; and P_2 is the probability that the event will result in a harm greater or equal to the collapse of civilization. For example, P_1 could represent the probability of nuclear war and P_2 could represent the probability that nuclear war would result in the collapse of civilization or worse. The occurrence of the initial catastrophe event does not necessarily entail the collapse of civilization—that depends on how effectively the survivors can cope with the aftermath of the event.

Calculating P_{EC} via Equation 1 requires two distinct analyses: one for each of P_1 and P_2. Analysis of P_1 is the analysis of the probability of initial events, and can follow many conventions of probabilistic risk analysis. In contrast, quantifying P_2 requires analysis of the severity, with attention to the success of catastrophe survivors. This is a rather different sort of analysis than is needed to quantify the probability of initial catastrophe events represented by P_1. However, P_2 is not equivalent to severity. P_2 is a probability variable representing the probability that the severity will exceed a certain threshold. P_2 can be obtained by creating a probability distribution for the severity of an initial event and then calculating the portion of that distribution that exceeds the threshold for existential catastrophe:

$$P_2 = \int_{S_T}^{\infty} P(S)\,\partial S \tag{2}$$

In Equation 2, P_2 is as in Equation 1; S is severity of some initial event; and S_T is the minimum severity threshold of existential catastrophe (the collapse of civilization in BRF). Equation 2 is illustrated in Figure 1.

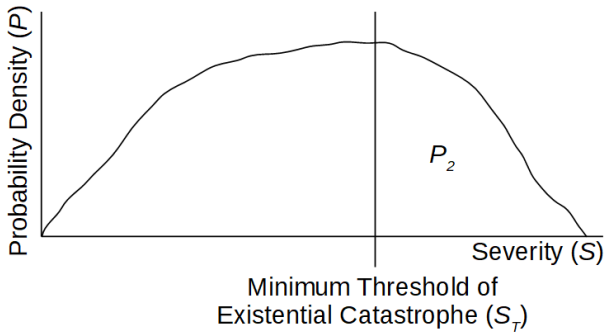

Figure 1. A depiction of the relationship between P_2 and the severity of an initial event. The graphic is for illustrative purposes only; no evaluation of the actual relationship is intended.

The studies surveyed in BRF vary in terms of which probabilities they quantify. (BRF acknowledges this; see BRF, Section 1.) Out of the 66 studies in BRF, only three—Rees (2003), Chapman (2004), and Wells (2009)—explicitly quantify the probability of civilization collapse (as obtained from a keyword search of BRF, Appendix A). Of the other 63 studies, some quantify P_1, while others quantify something similar to P_{EC} but with a different severity threshold.

Chapman (2004) is the only one of these studies that quantifies P_{EC} via something resembling Equation 1. Chapman (2004) is a study of asteroid impacts. As with many asteroid risk analyses, Chapman (2004) quantifies impact probability as a function of asteroid diameter (see Figure 1 of Chapman 2004). This provides P_1. Chapman (2004) then proposes that asteroids of 2-3km diameter may qualify as "civilization destroyers", while noting that the "predictions of consequences are fraught with uncertainty" (p.11; see also Baum 2018). This roughly amounts to postulating that $P_2 \approx 0$ for asteroids smaller than 2-3km diameter and $P_2 \approx 1$ for asteroids 2-3km diameter or larger. However, there is no reason to expect such an extreme discontinuity in P_2 at 2-3km. A more careful analysis would likely display a distribution of severities (as in Figure 1)

across a range of asteroid sizes, finding a continuum of P_2 values that gradually increase as a function of asteroid size. Until such analysis is conducted, one cannot rigorously quantify P_{EC} for asteroid impact. Given that impact probability varies widely as a function of diameter (smaller asteroids have a much higher impact probability), P_{EC} for asteroid impact will be highly sensitive to the resilience of civilization to impact events of different sizes. Similar reasoning applies to other existential risks.

The study of Wells (2009, p.84-90) is of note because it quantifies P_{EC} directly, without considering P_1 and P_2. It can do this because is quantifies the overall P_{EC}, not P_{EC} for specific events. The analysis uses a technique that quantifies the probability of something persisting based on how long it has already existed for. Given the current age of civilization, P_{EC} can be calculated. Wells (2009) also uses this technique to calculate the probability of human extinction, as do several other studies including Gott (1993). Because this technique does not consider any particulars of the existential risks that civilization currently faces, it provides at most a very limited amount of guidance to how the risks can be addressed.

The study of Rees (2003, p.8) is more ambiguous. Rees provides a subjective estimate of P_{EC}, stating "I think the odds are no better than fifty-fifty that our present civilization on Earth will survive to the end of the present century". The text does not explain how the estimate was made. The broader study presented by Rees (2003) includes reasoning similar to Wells (2009) and analysis of catastrophe scenarios similar to Chapman (2004). Perhaps the subjective estimate was made in consideration both approaches, but this cannot be discerned from the text.

The studies that quantify P_1 are not studies of the probability of existential catastrophe: they are studies of the probability of events that might or might not result in existential catastrophe. For example, Lipsitch and Inglesby (2014) quantify the probability that a highly pathogenic influenza strain would escape a research laboratory and cause a pandemic. Lipsitch and Inglesby (2014) also quantify the severity of the pandemic, estimating 2 million to 1.4 billion fatalities. To calculate P_2 and in turn P_{EC}, this range of severities would have to be evaluated in terms of the probability of exceeding S_T. Hellman (2008) quantifies the probability of a "full-scale nuclear war" started from a crisis similar to the Cuban missile crisis. Hellman (2008) does not quantify the severity and would need to do so to obtain P_2 and P_{EC}.

For the studies that quantify something analogous to P_{EC}, analysis of additional scenarios is needed to obtain P_{EC}. For example, Gott (1993) quantifies the probability of human extinction; additional analysis would quantify the probability of civilization collapsing without humanity going extinct. Pamlin and Armstrong (2015) quantify the probability of the subset of civilization collapse scenarios in which there is no subsequent recovery of civilization; additional analysis would quantify the probability of scenarios in which civilization collapses and recovers. Here it is worth recalling the multitude of definitions of existential catastrophe. The studies of Gott (1993) and Pamlin and Armstrong (2015) could classify as quantifying P_{EC} under non-BRF definitions.

2.3 Fast Vs. Slow Catastrophes

Existential risks can be classified in terms of how quickly the catastrophe unfolds. Specifically, it is the initial catastrophe event that unfolds at different speeds. The initial catastrophe event is the phenomenon that causes the harm, such as a nuclear war or Earth-asteroid collision; it is represented in P_1 in Equation 1. The overall catastrophe scenario also includes the aftermath of

the initial event. All existential catastrophes could have long-lasting effects, especially if civilization is not quickly rebuilt after it collapses. Nonetheless, the speed of initial events can vary significantly between different existential risks.

The risks included in BRF group into either fast or slow risks. Fast risks have initial events that occur within years or less. Fast risks included in BRF are nuclear war, pandemics, runaway artificial intelligence, asteroid impacts, nanotechnology, particle physics experiments, space weather, super-volcanic eruptions, and synthetic biology. Slow risks have initial events that persist for decades or longer. Slow risks included in BRF are climate change and ecological catastrophe (see BRF, Appendix A; note that the fast vs. slow distinction is my own and does not appear in BRF). This is overall a reasonable list of risks and broadly consistent with other studies of existential risk. The list is not without exceptions—for example, flood basalt volcanic eruptions consist of a series of large eruptions over a period of hundreds of thousands of years; if the eruptions are sufficiently large and frequent, they could cause more prolonged harm (Schmidt et al. 2016) and therefore rate as a slow catastrophe. As another example, studies of artificial intelligence sometimes consider that the event could unfold in a "slow takeoff" lasting on the order of "decades or centuries" (Bostrom 2014, p.63). Regardless, for probability analysis, the distinction between fast and slow is important in several respects. (The distinction can also be important for other reasons beyond the scope of this paper, such as policy decision-making.)

First, event probability P_1 can have a different meaning for slow risks. The probability of climate change is approximately 1 (i.e., it is virtually certain that climate change is occurring). The only meaningful uncertainty about climate change risk is about how severe the impacts will end up being, including whether they would amount to existential catastrophe. The same probably applies to the forms of ecological degradation that could result in ecological catastrophe, though the reference cited in BRF on ecological catastrophe (Pamlin and Armstrong 2015) is unclear on this point.

Second, slow catastrophes may be especially important when they interact with fast catastrophes. Climate change and ecological degradation may be unlikely to destroy civilization on their own. Instead, they could weaken civilization, making it less resilient to fast catastrophes. All else equal, climate change and ecological degradation may make it more likely that civilization survives a fast catastrophe (such as a nuclear war) if the fast catastrophe occurs in 2020 instead of 2050. (All else is not equal—other factors include, but are not limited to, economic and technological change.) In other words, climate change and ecological degradation increases P_2 for nuclear war (and other fast risks). They could also affect P_1, such as by destabilizing international politics, or alternatively by providing a rallying point that brings rival nations closer together. Additionally, fast catastrophes could also affect slow catastrophes. For example, the 2011 Fukushima Daiichi nuclear disaster, caused by the Tōhoku earthquake and tsunami, has prompted a shift from nuclear power to fossil fuels, thereby worsening climate change (Srinivasan and Rethinaraj 2013). While the Tōhoku-Fukushima event was only a local catastrophe, the episode is nonetheless indicative of potential impacts of fast catastrophes on slow catastrophes.

The above implies that the probability of slow catastrophe can be linked to the probability of fast catastrophe. This applies to most of the risks in BRF. One potential exception is particle physics experiments, for which P_1 may not be meaningfully affected by the slow catastrophes and for which P_2 may 1 under any circumstances (it may be impossible to survive a particle

physics catastrophe). Artificial intelligence could be another exception following the same reasoning, though it is conceivable that AI development could be altered by the desire to develop AI to address the slow catastrophes (Baum 2014), or by effects of slow catastrophes on the context in which AI is developed, such as related to international cooperation and competition. But for most cases, quantifying the probability of slow catastrophes should account for their effect on fast catastrophes.

It has been argued that the field of existential risk is insufficiently attentive to slow catastrophes, perhaps because slow catastrophes are too "boring" or "unsexy" (Liu et al. 2018; Kuhlemann 2019). My own view is that yes, slow catastrophes have been at least somewhat neglected in studies of existential risk. But it is also the case that slow catastrophes pose distinct analytical challenges, as outlined above. Furthermore, the fields dedicated to studying slow catastrophes tend to not study them in terms resembling existential risk. For example, high-profile research on planetary boundaries identifies approximately ten major global environmental threats that are nominally supposed to be unacceptable to humanity, but are defined in the research strictly in biogeophysical terms and not in terms of impacts to humanity (Baum and Handoh 2014). Likewise, almost all of the studies of slow catastrophes studied in BRF (Section 1) focus on biogeophysical metrics, in particular the magnitude of average global temperature increase from climate change. The one exception is Pamlin and Armstrong (2015), which is also probably the only of these studies that classifies within the field of existential risk. (BRF defines the existential risk research community as "those who are consciously seeking to align their research with the goal of understanding and managing such risks".) There is a major need for research that blends the traditions of global environmental change and existential risk to analyze, in human terms, the risks of slow catastrophes.

2.4 Multi-Risk Catastrophes

The preceding discussion of slow vs. fast catastrophes raises a more general point. The literature surveyed in BRF consists exclusively of quantifications of specific risks or quantification of the overall probability of existential catastrophe or similar catastrophe. However, some catastrophe scenarios involve multiple risks. These scenarios can require a different set of methods than those surveyed in BRF.

One type of scenario involves multiple initial catastrophe events occurring together by coincidence. For example, there could be a volcano eruption that happens to occur as a pandemic is breaking out. The initial events may be independent—for example, there is no reason to believe that infectious diseases affect plate tectonics and vice versa—in which case their probabilities (P_1 in Equation 1) could be modeled as independent random variables, with the aggregate initial event probability calculated accordingly. However, the probability of existential catastrophe given the co-occurrence of these events (P_2 in Equation 1) is unlikely to be independent. For example, a volcano eruption can prompt mass evacuations that facilitate disease transmission. Analysis of the severity of co-occurring initial catastrophe events must be done on a case-by-case basis; ditto for the accompanying probability of of existential catastrophe.

Another type of scenario involves causal relations between multiple fast catastrophes. For example, there could be a war or pandemic that causes a geoengineering termination shock (Baum et al. 2013; Parker and Irvine 2018), an asteroid collision that causes a nuclear war (Tagliaferri et al. 1994; Baum 2018), or a war that induces risky technology development (as

previously occurred in World War II with the development of nuclear weapons). These scenarios cannot be modeled as independent random variables. They can be modeled using fault trees or Bayesian networks, albeit with more complex models that account for the nuances of each catastrophe within the causal chain. In principle, they could be studied via aggregated opinion surveys, weighted aggregation, or enhanced solicitation, all of which derive from expert judgment. In practice, however, there may be a dearth of qualified experts, because experts tend to focus on one risk at a time and not on the causal relations between them. (The methods mentioned in this paragraph are described in Section 3.)

Finally, there are scenarios in which a fast catastrophe occurs during a slow catastrophe, as discussed in Section 2.3. Analysis of these scenarios could proceed by evaluating "baseline" values of P_1 and P_2 for the fast catastrophe, and then evaluating the effects of the slow catastrophe on the baseline P_1 and P_2. Alternatively, it could proceed by calculating scenario probabilities directly without first considering baselines. Regardless, the analysis requires a nuanced understanding of the impacts of the slow catastrophe as they relate to the onset and impacts of the fast catastrophe. (Other slow factors, such as economic and technological change, can be factored into the analysis alongside slow catastrophes.)

3. On the Methods Surveyed by BRF

BRF survey ten methods used for quantifying existential catastrophe:

1. Analytical approaches – estimating probability from theory without incorporating specific evidence on the risks, such as in the "doomsday argument".
2. Extrapolation and toy modeling – simple models that derive probabilities from data on other events on the assumption of an inherent relationship between the probability of the other events and existential catastrophe (noting that no existential catastrophe has previously occurred).
3. Fault trees – logic trees that decompose types of catastrophe events (e.g., nuclear war) into a collection of sequences of precursor events and conditions that can lead to the catastrophe event.
4. Bayesian networks – similar to fault trees, except also permitting interactions between precursor events (hence a "network" instead of a "tree").
5. Adapting large-scale models – the models depict details of specific risks (e.g., models of the global climate system to study climate change risk); these models are adapted to study extreme cases of these risks.
6. Individual subjective opinion – subjective judgments of a single person.
7. Aggregated opinion surveys – subjective judgments of multiple people, averaged into a single probability estimate.
8. Weighted aggregation – subjective judgments of multiple people, combined into a single probability estimate using a weighted average, with weights set according to some measure of the quality of the judgment expected from each person.
9. Enhanced solicitation – subjective judgments of multiple people, in which effort is made to improve judgment quality via some combination of training people to provide better opinions and being more careful in how their judgments are obtained.
10. Prediction markets – platforms for people to place financial bets on their best-guess parameter estimates.

3.1 The Tradeoff Between Quality and Accessibility

BRF evaluate probability quantification methods according to four criteria: (1) rigor, defined as how effectively the method makes use of available information; (2) uncertainty, defined as how effectively the method handles the considerable uncertainty about the probability of existential catastrophe; (3) accessibility, defined as how readily a research group can apply the method, especially the small interdisciplinary groups that often study existential risk; and (4) utility, defined as the value of the method's results for decision-making purposes.

BRF presents ratings of ten methods according to each of the four criteria. (The ratings were performed subjectively by the authors. In my own view, their ratings are reasonable.) Ideally, a method would rate well across all four criteria. Unfortunately, none of the methods surveyed by BRF achieve this.

The BRF ratings data show a trend in which rigor, uncertainty, and utility are similar to each other and dissimilar to accessibility. This trend holds across all methods (see BRF, Table 1). To illustrate this trend, Figure 2 plots rigor vs. accessibility. No method is rated high in both, and only weighted aggregation is rated low in both. (More precisely, weighted aggregation is rated medium-low in rigor.)

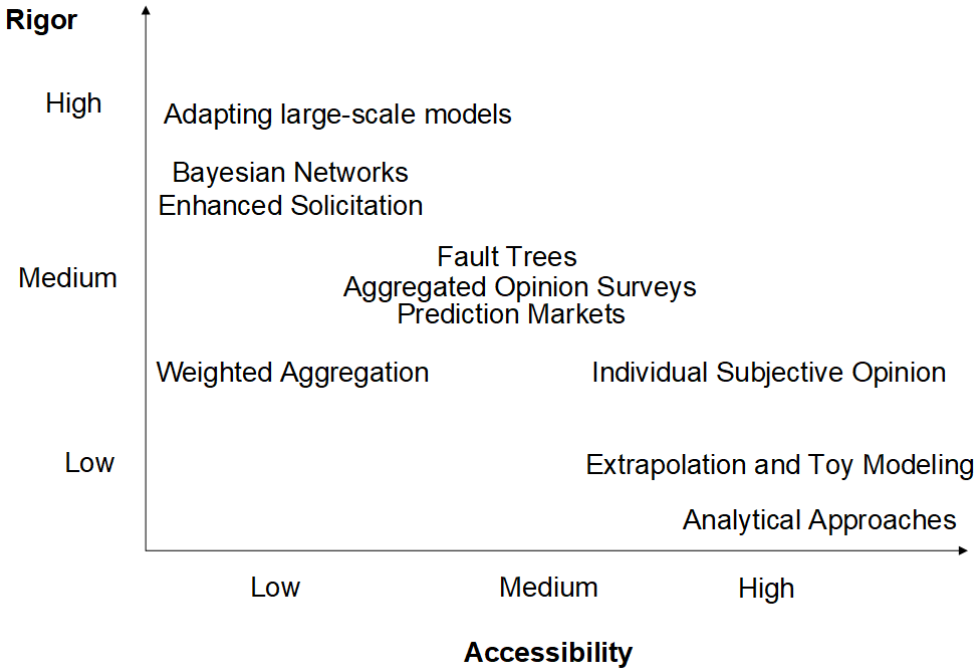

Figure 2. Rigor vs. accessibility of each method as rated by BRF (data from BRF, Table 1).

The trend in the data is entirely understandable. Rigor, uncertainty, and utility are all aspects of the quality of the probability quantification produced by the application of a particular method. Obtaining higher-quality results generally requires incorporating more detail about the risk and more nuance in how the detail is processed, both of which tend to make the method less accessible. There can be methods that are low-quality and inaccessible, though one would hope that these have gotten weeded out of the collective risk analysis toolkit. Conversely, however, it

may not be possible for a method to be both high-quality and accessible. Quality analysis may inevitably require a substantial investment.

The same point applies to applications of a given method. Any method can be applied in ways that are of higher or lower quality. There can be applications of a given method that are low-quality and inaccessible: a lot of resources spent implementing a method poorly. Higher quality implementations of a given method will tend to be less accessible, requiring more effort on the part of the research team, and in some cases more funding to invest in expenses such as travel to interview experts in person (instead of via phone or other mode of remote communication) or rewards to incentivize participants in prediction markets.

4. On the Role of Existential Risk Quantification

4.1 Why Quantify Probability (and Severity)?

Underlying the tradeoff between quality and accessibility is the basic point that the probabilities of potential existential catastrophe scenarios are difficult to rigorously quantify. The scenarios involve complex and unprecedented global processes. Historical data are unreliable or nonexistent. Expert judgment is often unreliable as well—or, perhaps there are no experts. Obtaining quality quantifications requires applying difficult methods, which can be a resource-intensive process in a field in which resources are scarce. Perhaps it would be better to skip the matter entirely and focus on other activities.

BRF identify two major reasons to quantify the probabilities and severities of existential risks. The first is the prioritization of efforts to reduce various risks. The second is the evaluation of actions that could decrease one set of risks but increase another (i.e., risk-risk tradeoffs). Both are situations in which actions could affect multiple risks, and in which identifying the best action requires comparing the size of the effect on one set of risks to the size of the effect on another. These sizes are typically measured in terms of the product of probability and severity. Therefore, to identify the best actions in these situations, it is necessary to know the effects of the actions on the probabilities and severities of the risks in question. Quantifying the probabilities and severities is an important first step. (Additional analysis is required to assess the effects of specific actions aimed at reducing the risks, as well as the costs associated with pursuing these actions.)

Quantification is not always necessary. Some actions would only affect one risk, or would affect multiple risks in the same direction (increase or decrease). These actions pose no tradeoff and require no quantification to evaluate. Where possible, it can be more efficient to focus on these actions. Indeed, one approach to handling difficult quantification tradeoffs is to search for actions that avoid the tradeoffs. Graham and Wiener (1995) refer to these as “risk-superior” options, analogous to the concept of Pareto superiority in welfare economics. An “existential-risk-superior” option would be one that reduces all existential risks by at least as much as all other options, and reduces at least one existential risk by more than each other option, such that there is reason to favor the existential-risk-superior option and no reason to favor any other option, setting aside reasons unrelated to existential risk. If existential-risk-superior options are identified, there is no further need to quantify the options, because one option is clearly best. However, even then, some quantification may be needed to identify existential-risk-superior

options. And in many other cases, there are no existential-risk-superior options. Tradeoffs are inevitable. Quantification is an important research task, however difficult it may be.

4.2 Analysis vs. Decision-Making

Some further perspective can be obtained from considering the relationship between analysis and decision-making.

Arguably, analysis of the probability and severity of existential catastrophe should aim to improve the quality of decisions that affect existential risk. Otherwise, the analysis is of mere intellectual interest. Reducing existential risk is (again, arguably) a moral imperative, not an intellectual curiosity. Analysis can improve decisions by enabling the decisions to result in larger reductions in the risk. Phrased differently, the expected value of the action taken is larger with the analysis than without. This is the essence of the concept of the value of information (e.g., Barrett 2017). Analysis can be valuable even if the action taken is unchanged. As a toy example, perhaps heads would have been chosen in a coin flip anyway, but it would have been a lucky guess, and so there is value in analysis indicating that the coin was indeed going to land on heads. The same reasoning holds for analysis about more complex decisions, including those about existential risk.

But decisions do not occur in a theoretical vacuum, and they do not necessarily conform to the ideals of expected value maximization. Decisions are psychological and (often, especially for the most important decisions) social processes. In order to be most useful for reducing the risks, methodology for quantifying probabilities and severities should be designed with the nuances of actual decision processes in mind. This point is captured somewhat by the BRF concept of method utility, but some further elaboration is warranted.

A lot of risk analysis is commissioned by clients who seek to use the analysis in their own decision-making. The field of risk analysis has roots in industry, especially nuclear power (Thompson et al. 2005; Wilson 2012). Engineering risk analysis commonly seeks to guide in-house safety decisions of industrial risk managers or regulatory decisions by policy makers at agencies like the US Environmental Protection Agency and Nuclear Regulatory Commission. For this, quantifying the overall probability of some catastrophe can be less important than identifying the specific system components that are most implicated in the risk and likewise in greatest need of attention. This is one reason that engineering risk analysis has gravitated toward methods like fault trees that decompose risks into constituent parts and therefore produce results that help managers focus on specific problems.

Existential risk research is less often commissioned by clients. Instead, it is often sponsored philanthropically, whether by private funders or government agencies. The funders do not seek results that they can use for their own decisions. Instead, they wish to support the cause of existential risk reduction and hope that the analysis will in some way be useful to that end. Without decision-maker clients, existential risk analysis projects must take additional steps to ensure that their analysis improves decision-making, including identifying decision-makers to support. Because of this, there is a greater risk of the analysis working in an academic fashion, disconnected from practical decision-making and not having constructive impact on the risks themselves. (Existential risk is not the only domain in which research lacks decision-maker clients. Another example is research by industry and environmental public interest groups that seek to inform environmental policy; von Winterfeldt et al. 2012.)

Nonetheless, even without built-in decision-maker clients, it is still important for existential risk analysts to bear in mind how their analysis could be used for decision-making. That means recognizing that there is more to risk management than just numbers for probability and severity. The details of the risks themselves must be understood, and how potential actions can affect the risks, and the psychological and social and institutional contexts in which decisions will be made. This is a lot to account for—as if the probabilities and severities were not a lot on their own—but it is essential for pursuing actual reductions in existential risk and not just academic studies.

It can help to recognize that risk analysis can be valuable even if it does not produce any quantification of probability and severity. Often, simply having an analysis that pulls together information about the risk in an intuitive and well-organized fashion is sufficient for improving risk management. In other words, risk analysis can be of value by improving people's mental models of risks, in addition to providing information on the rating or ranking of decision options. Additionally, risk analysis can be of value by creating an opportunity for stakeholders and experts to share risk-related information with each other in a way that crosses typical organizational barriers. The opportunity for people in different groups to talk through key issues affecting risk can be of significant valuable for the practice of risk management, even regardless of formal risk analysis outputs. It is important to not let the pursuit of numbers (of probability, severity, the timing of future events such as AI milestones, etc.) serve as a distraction from the goals of evaluating decision options and improving our understanding and management of the risks.

In sum, quantifying probability and severity should be at most only one part of an overall existential risk analysis and management portfolio. But it is still a part, and an important one at that. It is likewise important to use quality methodology in performing the quantifications.

5. Conclusion

This article has sought to build on the excellent contribution of BRF to provide further perspective on the quantification of the probability of existential catastrophe. This article finds that the probability of existential catastrophe is inherently linked to the severity of events that could result in existential catastrophe, that achieving a higher quality of analysis of the probability will in general require a larger investment in analysis, that analysis of the probability is sometimes but not always necessary for decision-making, and that analysis should be designed to support risk management and not just designed as an academic exercise. If these findings are taken into account, together with the analyses of specific methodologies by BRF, then it is believed that risk analyses of existential catastrophe will tend to be more successful at understanding and reducing the risks.

Acknowledgments

Tony Barrett, Seán Ó hÉigeartaigh, Robert de Neufville, Simon Beard, James Fox, David Manheim, Siebe Rozendal, two anonymous reviewers, and editor Ted Fuller provided helpful feedback on an earlier version. Any remaining errors are the author's alone.

References

Barrett, A. M. (2017). Value of global catastrophic risk (GCR) information: Cost-effectiveness-based approach for GCR reduction. *Decision Analysis*, 14(3), 187-203.

Baum, S. D. (2014). The great downside dilemma for risky emerging technologies. *Physica Scripta*, 89(12), 128004.

Baum, S. D. (2018). Uncertain human consequences in asteroid risk analysis and the global catastrophe threshold. *Natural Hazards*, 94(2), 759-775.

Baum, S. D., & Handoh, I. C. (2014). Integrating the planetary boundaries and global catastrophic risk paradigms. *Ecological Economics*, 107, 13-21.

Baum, S. D., Maher, T. M., & Haqq-Misra, J. (2013). Double catastrophe: Intermittent stratospheric geoengineering induced by societal collapse. *Environment Systems & Decisions*, 33(1), 168-180.

Baum, S. D., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., et al. (2019). Long-term trajectories of human civilization. *Foresight*, 21(1), 53-83.

Beard, S., Rowe, T., & Fox, J. (2020). An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards. *Futures*, 115, 102469

Bostrom, N., (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9(1).

Bostrom, N., (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford, U.K.: Oxford University Press.

Chapman, C. R. (2004). The hazard of near-Earth asteroid impacts on earth. *Earth and Planetary Science Letters*, 222(1), 1-15.

Graham, J. D., & Wiener, J. B. (1995). *Risk vs. Risk: Tradeoffs in Protecting Health and the Environment*. Cambridge, M.A.: Harvard University Press.

Gott, J. R. III (1993). Implications of the Copernican principle for our future prospects. *Nature*, 363, 315-319.

Grossi, P., Kunreuther, H., & Windeler, D. (2005). An introduction to catastrophe models and insurance. In Grossi, P., & Kunreuther, H. (Eds.), *Catastrophe Modeling: A New Approach to Managing Risk*. Boston: Springer, pp. 23-42.

Hellman, M. (2008). Risk analysis of nuclear deterrence. *The Bent of Tau Beta Pi*, 99(2), 14-22.

Kuhlemann, K. (2019). Complexity, creeping normalcy and conceit: Sexy and unsexy catastrophic risks. *Foresight*, 21(1), 35-52.

Lipsitch, M., & Inglesby, T. V. (2014). Moratorium on research intended to create novel potential pandemic pathogens. *Mbi*o, 5(6), e02366-14, doi:10.1128/mBio.02366-14.

Liu, H. Y., Lauta, K. C., & Maas, M. M. (2018). Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures*, 102, 6-19.

Morrison, D. (1992). *The Spaceguard Survey: Report of the NASA International Near-Earth-Object Detection Workshop*. Washington, D.C.: National Aeronautics and Space Administration.

Pamlin, D. & Armstrong, S. (2015). *Global Challenges: 12 Risks that Threaten Human Civilisation*. Stockholm: Global Challenges Foundation.

Parker, A., & Irvine, P. J. (2018). The risk of termination shock from solar geoengineering. *Earth's Future*, 6(3), 456-467.

Persson, I. & Savulescu, J. (2012). *Unfit for the Future: The Need for Moral Enhancement*. Oxford: Oxford University Press.

Rees, M. J. (2003). *Our Final Century: Will the Human Race Survive the Twenty-First Century?* London: Heinemann.

Schmidt, A., Skeffington, R. A., Thordarson, T., Self, S., Forster, P. M., Rap, A., et al. (2016). Selective environmental stress from sulphur emitted by continental flood basalt eruptions. *Nature Geoscience*, 9(1), 77-82.

Sotala, K., & Gloor, L. (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica*, 41(4) 389-400.

Srinivasan, T. N., & Rethinaraj, T. G. (2013). Fukushima and thereafter: Reassessment of risks of nuclear power. *Energy Policy*, 52, 726-736.

Tagliaferri, E., Spalding, R., Jacobs, C., Worden, S. P., & Erlich, A. (1994). Detection of meteoroid impacts by optical sensors in Earth orbit. In Gehrels, T., Matthews, M. S., & Schumann, A. (Eds.), *Hazards Due to Comets and Asteroids*. Tucson: University of Arizona Press, pp.199-221

Thompson, K. M., Deisler Jr, P. F., & Schwing, R. C. (2005). Interdisciplinary vision: The first 25 years of the Society for Risk Analysis (SRA), 1980–2005. *Risk Analysis*, 25(6), 1333-1386.

Tonn, B. E. (1999). Transcending oblivion. *Futures*, 31, 351-359.

Tonn, B., & Stiefel, D. (2013). Evaluating methods for estimating existential risks. *Risk Analysis*, 33(10), 1772-1787.

von Winterfeldt, D., Kavet, R., Peck, S., Mohan, M., & Hazen, G. (2012). The value of environmental information without control of subsequent decisions. *Risk Analysis*, 32(12), 2113-2132.

Wells, W. (2009). *Apocalypse When? Calculating How Long the Human Race Will Survive*. Chichester, U.K.: Praxis.

Wilson, R. (2012). The development of risk analysis: A personal perspective. *Risk Analysis*, 32(12), 2010-2019.